BIG DATA & FINANCIAL TECHNOLOGY INTRO

SETH H. HUANG

Seth H. Huang, PhD

PhD Cornell University
MA Cornell University
BA Boston University

Trader/ fund manager for financial institutions and hedge funds

BEFORE DELVING INTO FINTECH, WE MUST KNOW WHAT'S CHANGING TODAY.

A FEW WORDS

Big data is a hot topic, but it is not a fad. It has changed how most corporations conduct businesses. It has changed how people analyze data worldwide.

It is mostly done by engineers, but their knowledge and method can be contained by using existing tools.

The estimate is that by 2020, we will only have 50% of engineers filling for big-data-related positions. If you can understand and acquire a basic skill set, it can serve you well in different fields.

However…

This topic is vast. We can only do a very general view today.

"what big data is"

"how people use it"

"what do people use"

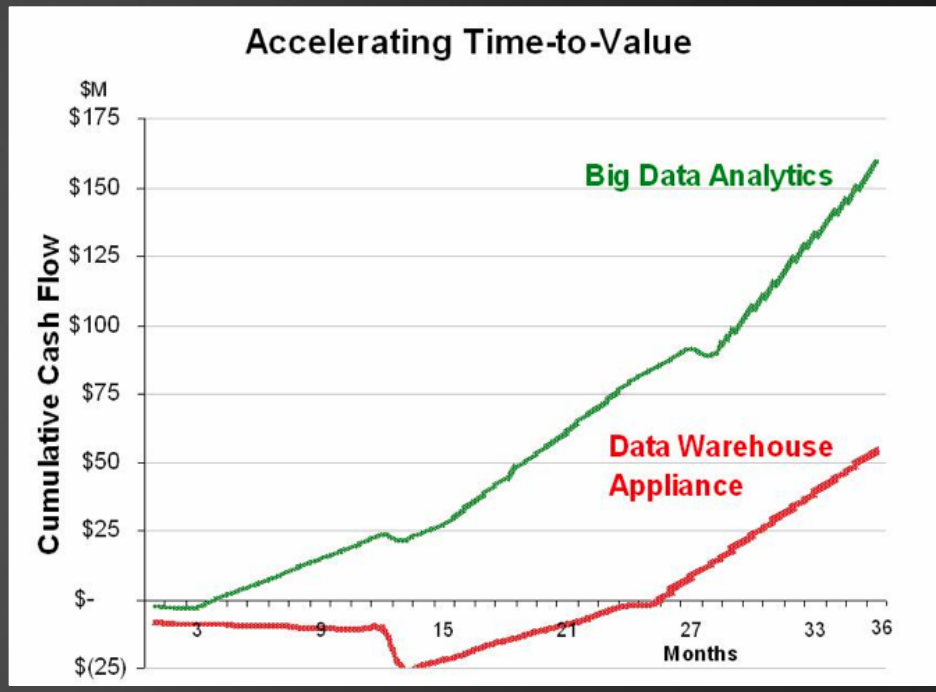
Which leads us to:

WHAT IS BIG DATA?

What is Big Data?
What makes data “Big” Data?

Also called Big Data Analytic

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



WHY BIG DATA?

Essentially, the whole field is developed to manage very large amounts of data and extract value and knowledge from them

BIG DATA DEFINITION…

- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

OR

Big data is the scale of data that is so large such that most computer memories cannot hold, process and analyze it.

What's Big Data?

- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization. With the last time being the whole value-driver of big data.
- The trend to larger data sets is due to:

The additional information from the analysis of a single large data set, as compared to separate smaller sets with the same total amount of data, helps analysts "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

According to McKinsey, companies who miss big data opportunities of today will miss the next frontier of innovation, competition, and productivity.

But the idea of big data is still very vague for a majority of populations, and only large corporations are tapping into it.

BIG DATA 3Vs

There are usually three data categories:

- **Volume**
- **Variety**
- **Velocity**

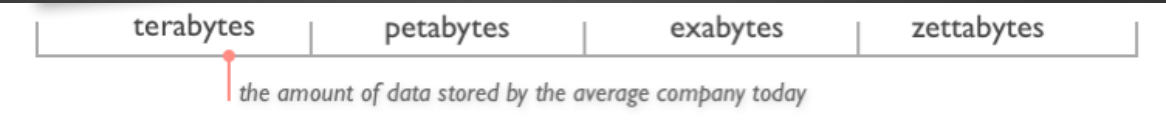
Unlike most corporate jargons, the 3Vs in big data actually ARE IMPORTANT, as it describes the capacity required and the type of data.

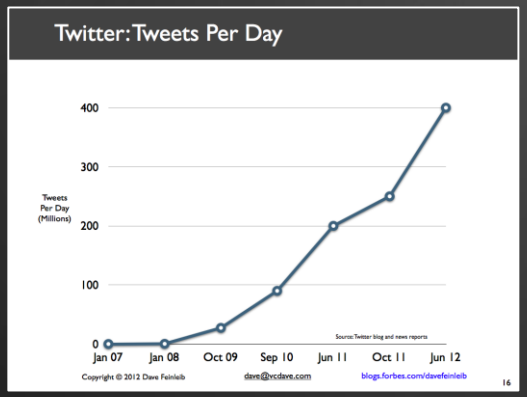
Characteristics of Big Data: 1-Scale (Volume)

Data Volume

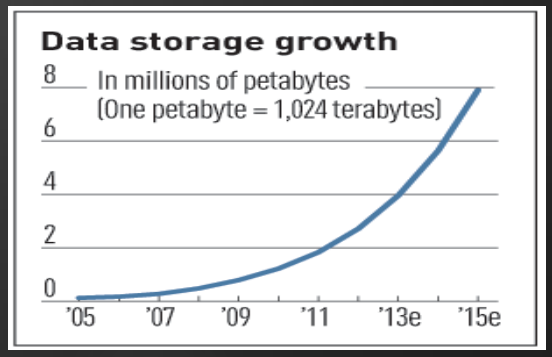
- 44x increase from 2009 to 2020 (est.)
- From 0.8 zettabytes to 35zb

Data volume is increasing exponentially





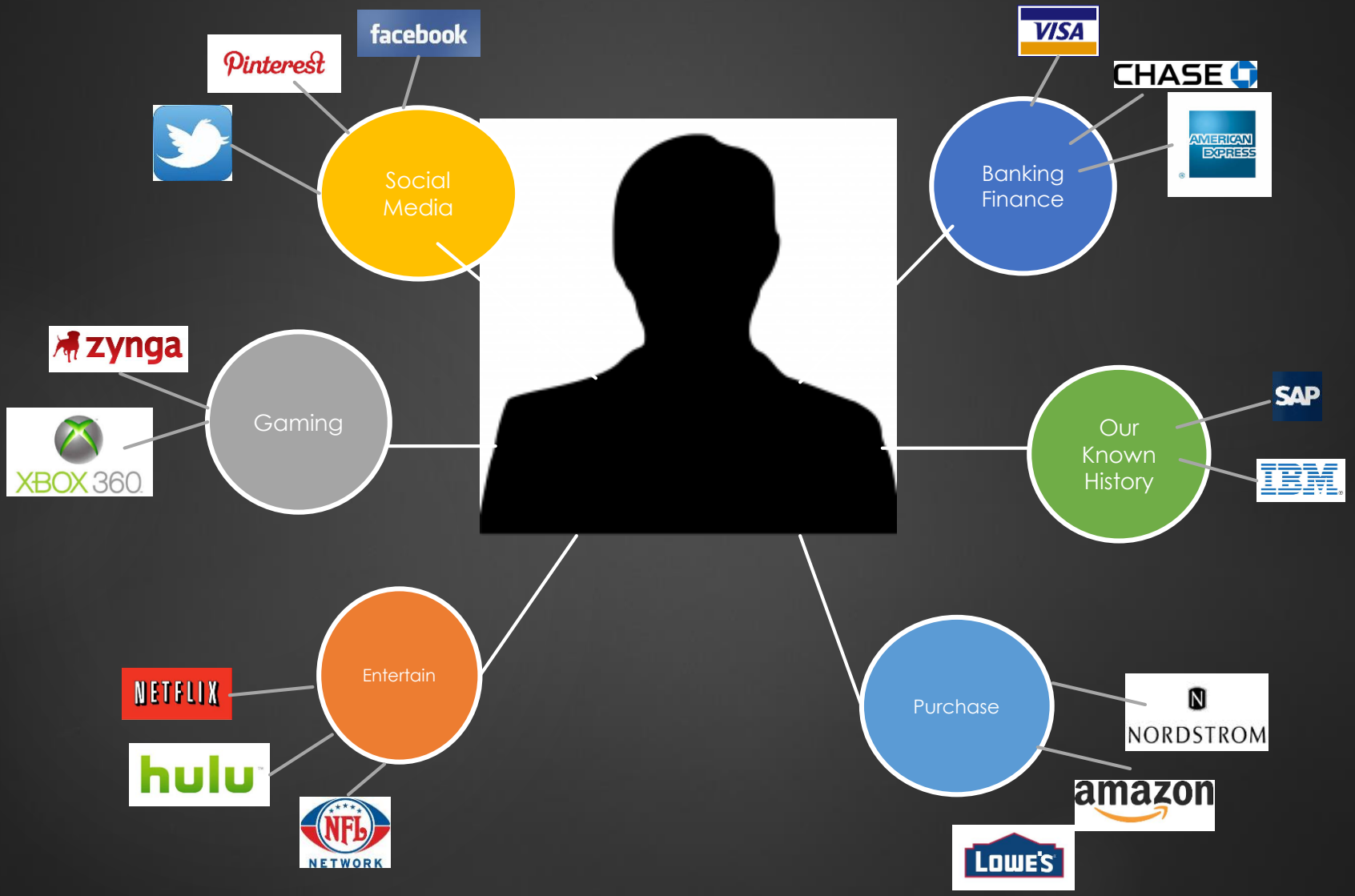
Exponential increase in collected/generated data



Characteristics of Big Data: 2-Complexity (Varity)

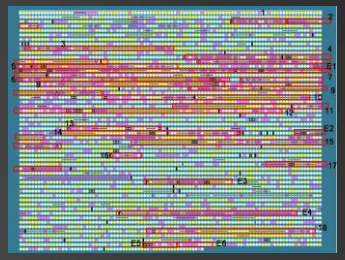
First, just by yourself, imagine what kinds of data you generate every day.

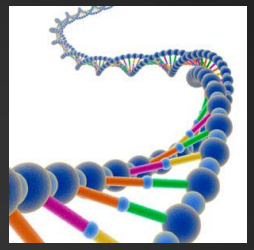
A Single View to the Customer

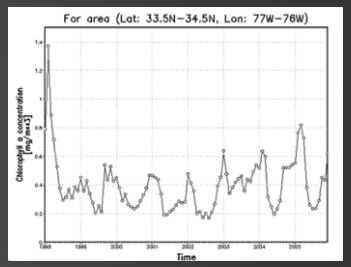


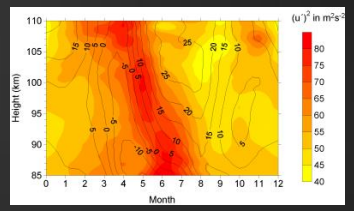
Characteristics of Big Data: 2-Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data





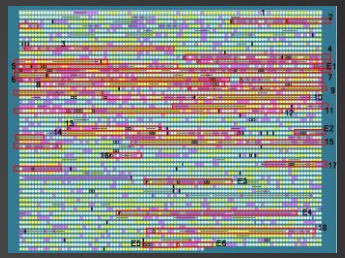


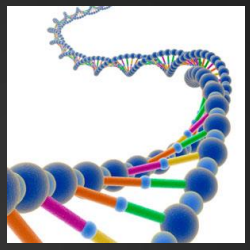


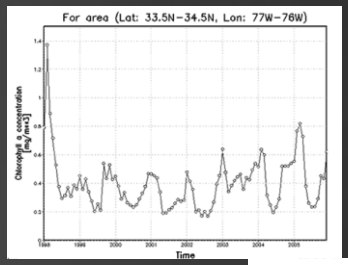
To extract knowledge➔ all these types of data need to linked together

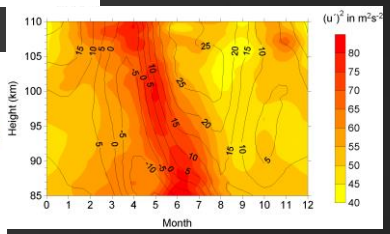
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)









To extract knowledge➔ all these types of data need to linked together

Characteristics of Big Data: 3-Speed (Velocity)

Data is begin generated fast and need to be processed fast

"Online" Data Analytics

Late decisions ➔ missing opportunities

Examples

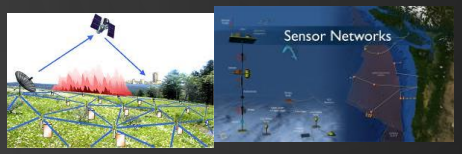
- **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
- **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

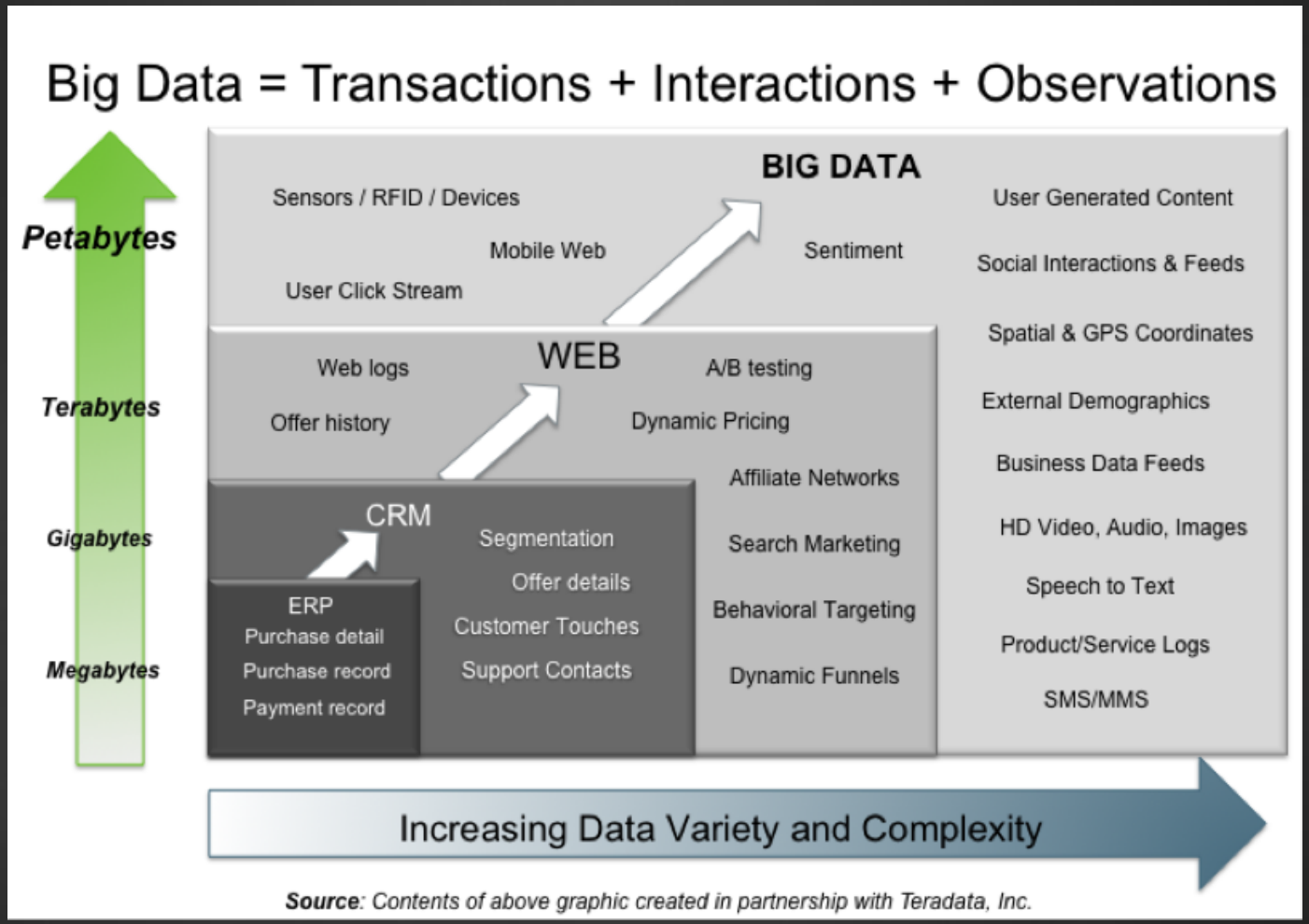
Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

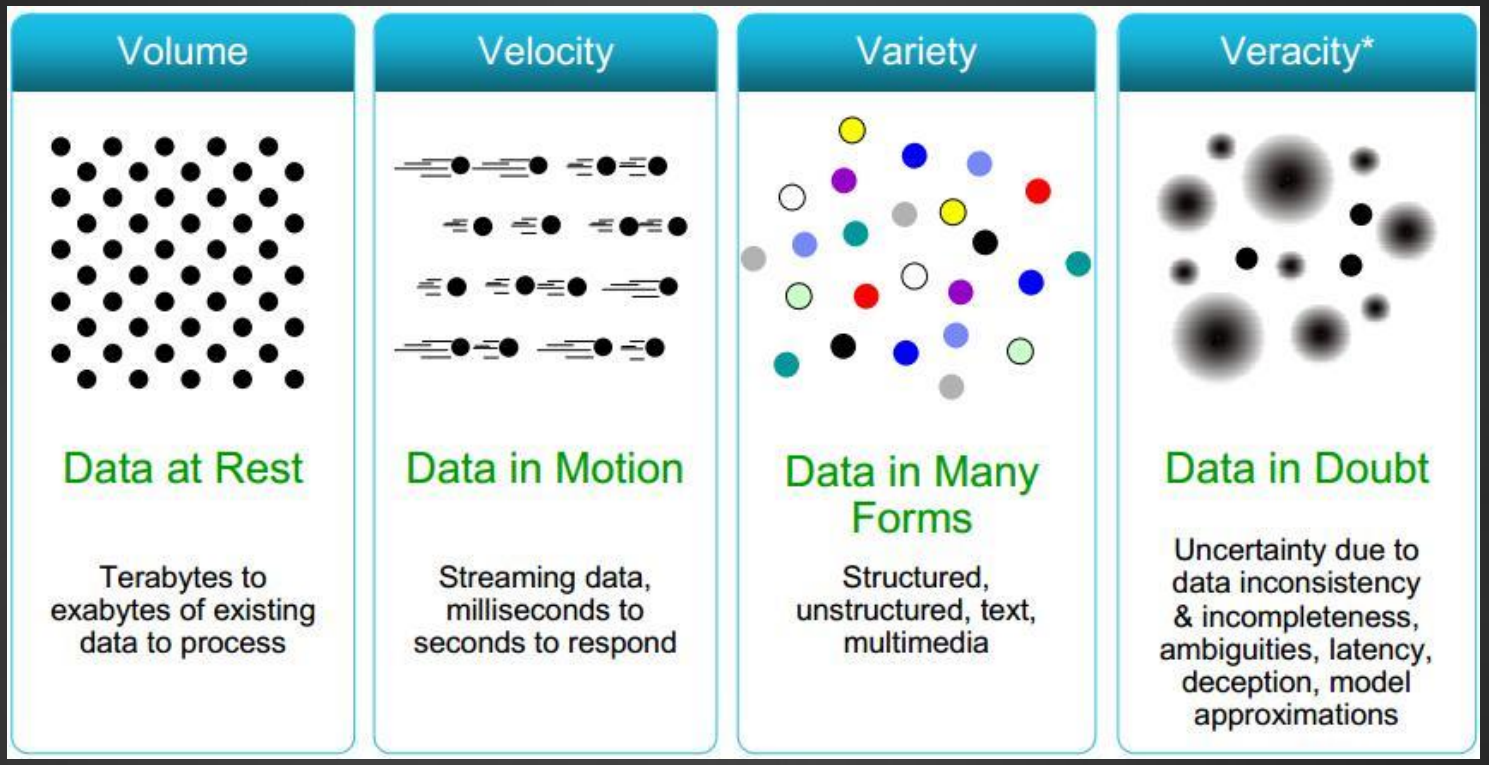
- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Big Data: 3V's



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Some Make it 4V's




The Future of Robotics and Artificial Intelligence (Andrew Ng, Stanford Unive...

WHAT COMPANIES ARE IN BIG DATA INFRASTRUCTURE

WHAT TECHNOLOGIES/ SKILLS DO PEOPLE NEED TO PERFORM BIG DATA ANALYTIC?

(HINT: YOU DON'T HAVE TO BE A ROCKET SCIENTIST, BUT…)

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence

ORACLE | Hyperion
SAP Business Objects RJMetrics
Microsoft | Business Intelligence
IBM COGNOS birst
MicroStrategy
Autonomy
bime
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA
metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS TIBCO
KARMASPHERE
panopticon
Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA

Data As A Service

factual
kaggle
knoema beta
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
LOQATE

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
ParAccel
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

CouchBase
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

Technologies

hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

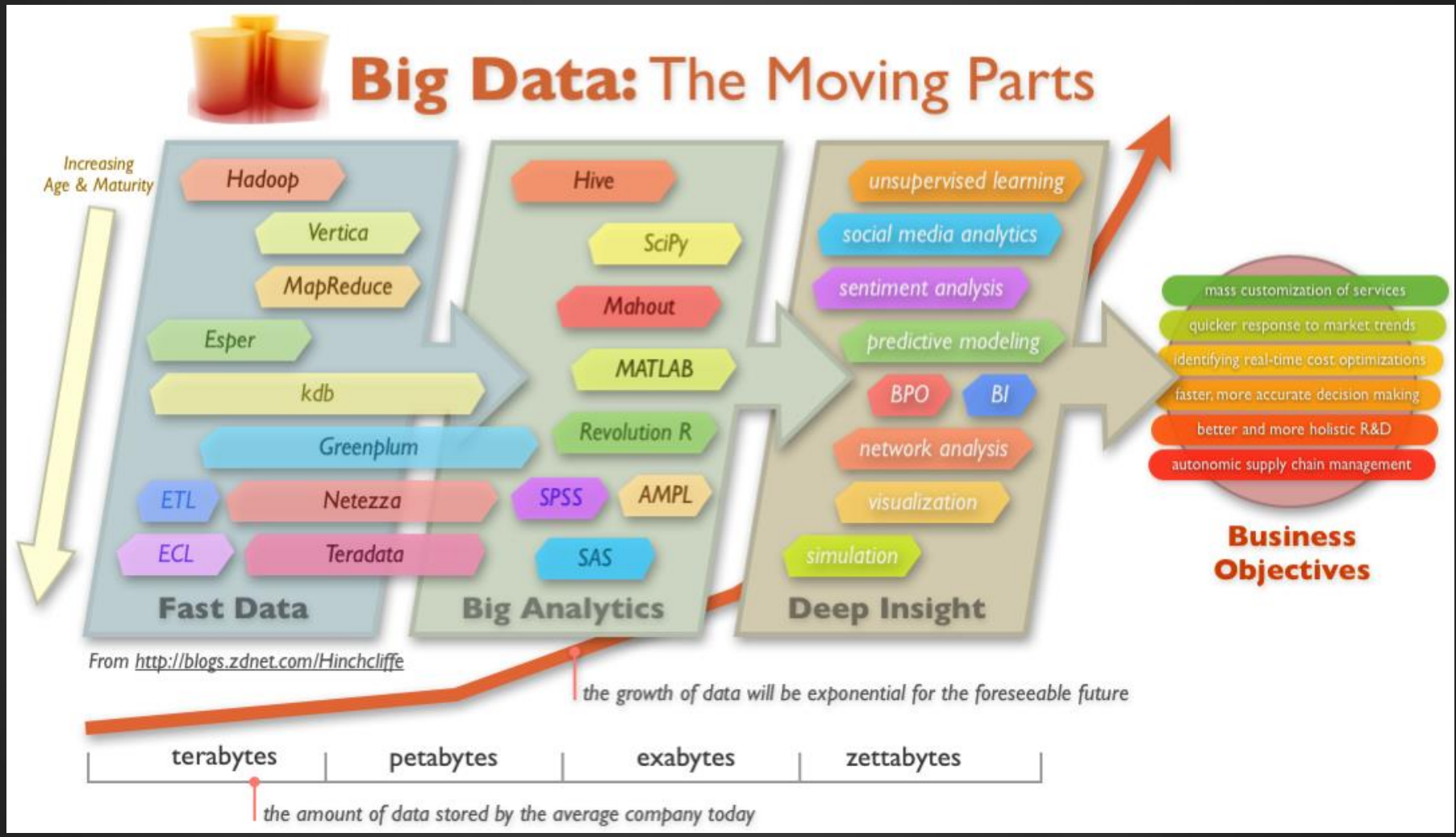
dave@vcdave.com
blogs.forbes.com/davefeinleib

BASICALLY, THESE PEOPLE KNOW:

Hadoop/MapReduce technology

- **Learn the platform (how it is designed and works)**
  - How big data are managed in a scalable, efficient way
- **Learn writing Hadoop jobs in different languages**
  - Programming Languages: Java, C, Python
  - High-Level Languages: Apache Pig, Hive
- **Learn advanced analytics tools on top of Hadoop**
  - RHadoop: Statistical tools for managing big data
  - Mahout: Data mining and machine learning tools over big data
- **Learn state-of-art technology from recent research papers**
  - Optimizations, indexing techniques, and other extensions to Hadoop

Big Data Technology: ALSO OUR FOCUS



ANOTHER BUZZWORD WAS "CLOUD COMPUTING," A GOOD BUDDY WITH "BIG DATA"

MANY THINGS NOW ARE DONE ON CLOUDS BECAUSE NO SINGLE COMPUTER CAN HANDLE THE AMOUNT OF DATA.

Cloud Computing

IT resources provided as a service:

- Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
- Cheap storage, high bandwidth networks & multicore processors
- Geographically distributed data centers
- Offerings from Microsoft, Amazon, Google, …

Benefits

Cost & management

- Economies of scale, “out-sourced” resource management

Reduced Time to deployment

- Ease of assembly, works “out of the box”

Scaling

- On demand provisioning, co-locate data and compute

Reliability

- Massive, redundant, shared resources

Sustainability

- Hardware not owned

Classification of Cloud Computing based on Service Provided

- Infrastructure as a service (IaaS)
  - Offering hardware related services using the principles of cloud computing. These could include storage services (database or disk storage) or virtual servers.
  - Amazon EC2, Amazon S3, Rackspace Cloud Servers and Flexiscale.
- Platform as a Service (PaaS)
  - Offering a development platform on the cloud.
  - Google's Application Engine, Microsofts Azure, Salesforce.com's force.com .
- Software as a service (SaaS)
  - Including a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector.
  - Salesforce.coms' offering in the online Customer Relationship Management (CRM) space, Googles gmail and Microsofts hotmail, Google docs.

Refined Categorization

- Storage-as-a-service
- Database-as-a-service
- Information-as-a-service
- Process-as-a-service
- Application-as-a-service
- Platform-as-a-service
- Integration-as-a-service
- Security-as-a-service
- Management/ Governance-as-a-service
- Testing-as-a-service
- Infrastructure-as-a-service

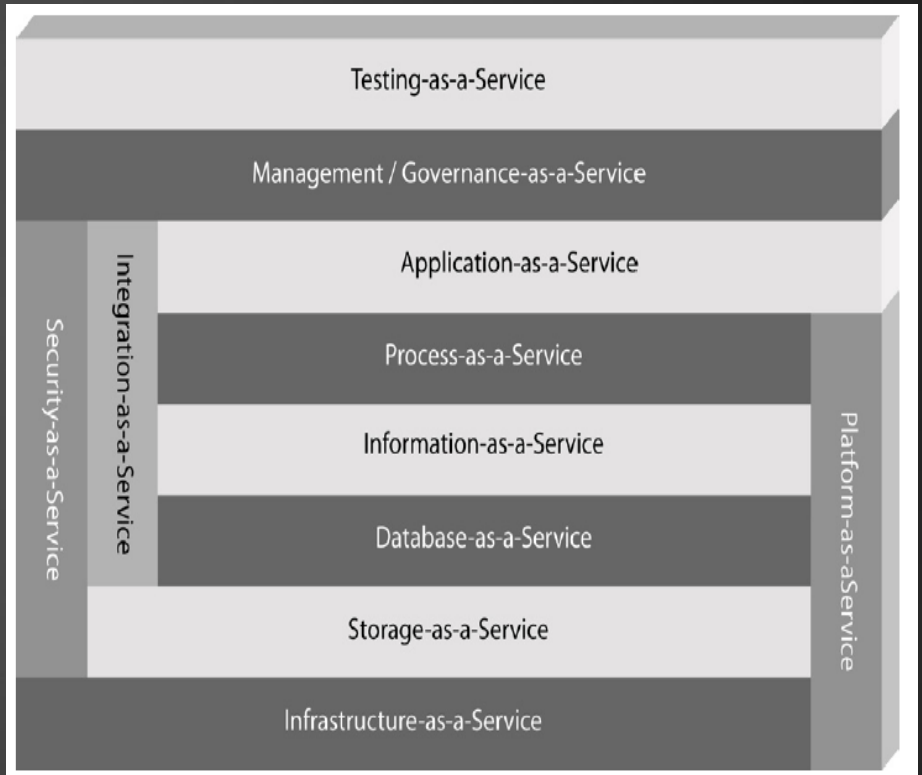


Figure 1: The patterns or categories of cloud computing providers allow you to use a discrete set of services within your architecture.

InfoWorld Cloud Computing Deep Dive

Everything as a Service

- Utility computing = Infrastructure as a Service (IaaS)
  - Why buy machines when you can rent cycles?
  - Examples: Amazon's EC2, Rackspace
- Platform as a Service (PaaS)
  - Give me nice API and take care of the maintenance, upgrades, …
  - Example: Google App Engine
- Software as a Service (SaaS)
  - Just run it for me!
  - Example: Gmail, Salesforce

AMAZON WEB SERVICES (AWS)

- Elastic Compute Cloud – EC2 (IaaS)
- Simple Storage Service – S3 (IaaS)
- Elastic Block Storage – EBS (IaaS)
- SimpleDB (SDB) (PaaS)
- Simple Queue Service – SQS (PaaS)
- CloudFront (S3 based Content Delivery Network – PaaS)
- Consistent AWS Web Services API

To know1: Data Analytics & Data Mining

- Exploratory Data Analysis
- Linear Classification (Perceptron & Logistic Regression)
- Linear Regression
- C4.5 Decision Tree
- Apriori
- K-means Clustering
- EM Algorithm
- PageRank & HITS
- Collaborative Filtering

To know 2: Hadoop/MapReduce Programming & Data Processing

- Architecture of Hadoop, HDFS, and Yarn
- Programming on Hadoop

- Basic Data Processing: Sort and Join
- Information Retrieval using Hadoop
- Data Mining using Hadoop (Kmeans+Histograms)
- Machine Learning on Hadoop (EM)

- Hive/Pig
- HBase and Cassandra

FOR THIS COURSE, WE WILL JUST LOOK THE "PREDICTIVE PART" OF BIG DATA. SPECIFICALLY, LET'S WATCH THIS VIDEO.

THE RECENT BIG DATA DEVELOPMENT

The Business Analytic Model Has Changed. Forever.

- **The Model of Generating/Consuming Data has Changed**

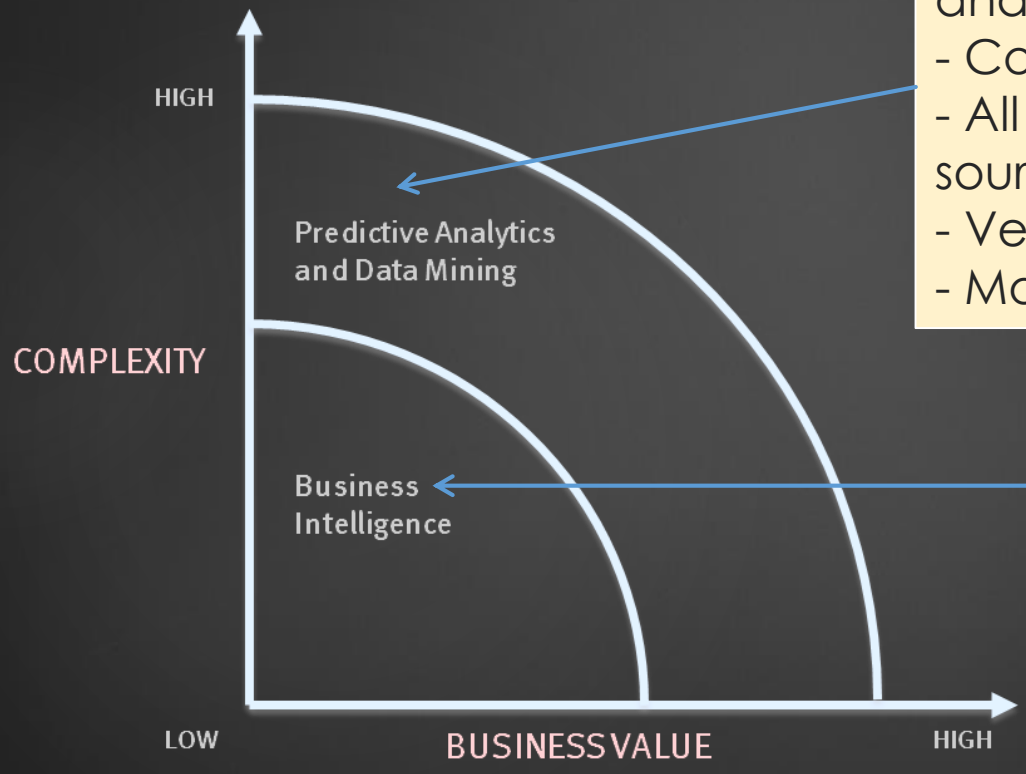
Old Model: Static and small. Few companies are generating data, all others are consuming data



New Model: Dynamic and vast. all of us are generating data, and all of us are consuming data



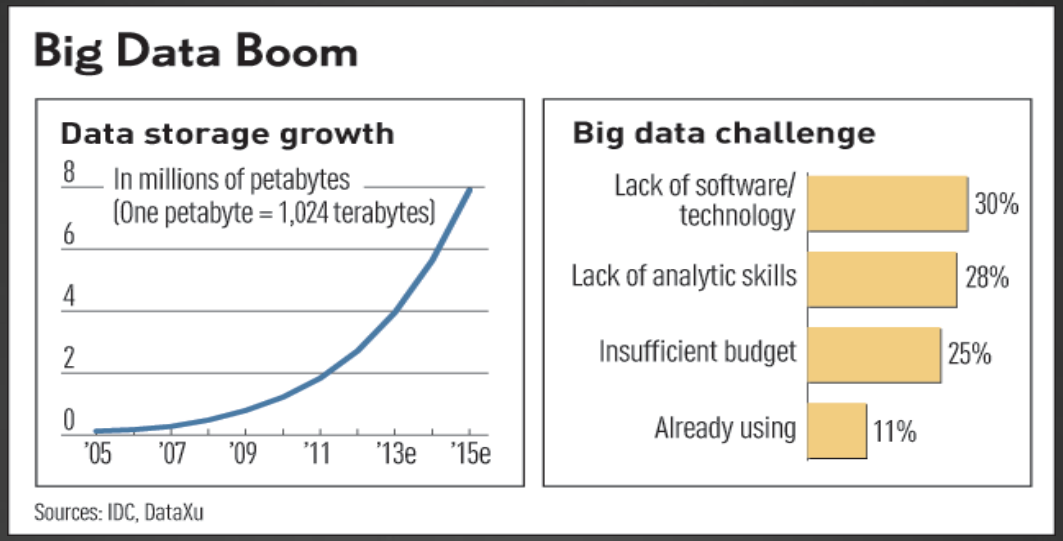
What's driving Big Data as an industry?



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

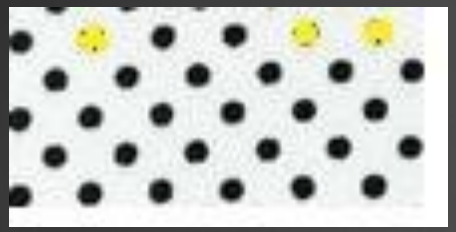
Challenges in Handling Big Data

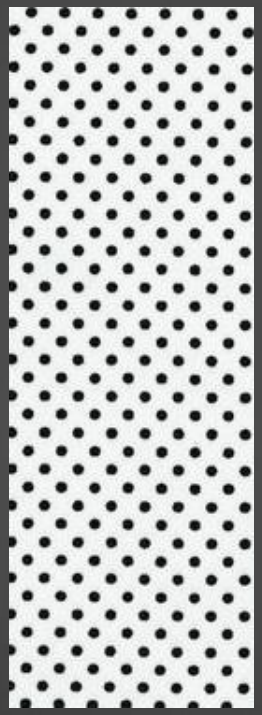


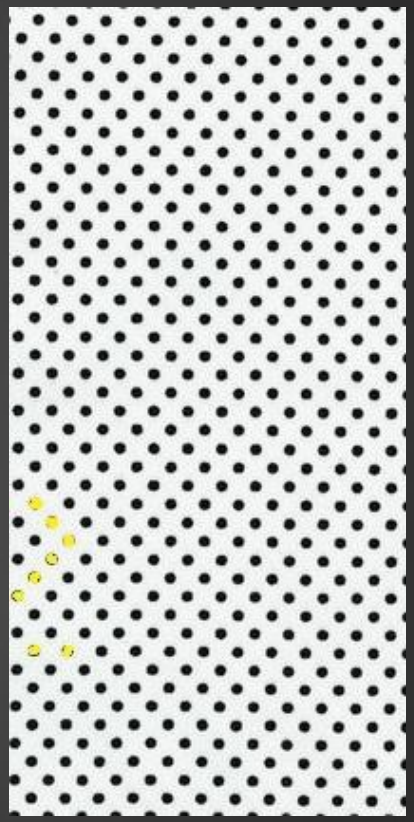
- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

Finding the right information can be difficult. Can be misleading, even.

Data is often pieces of puzzle.

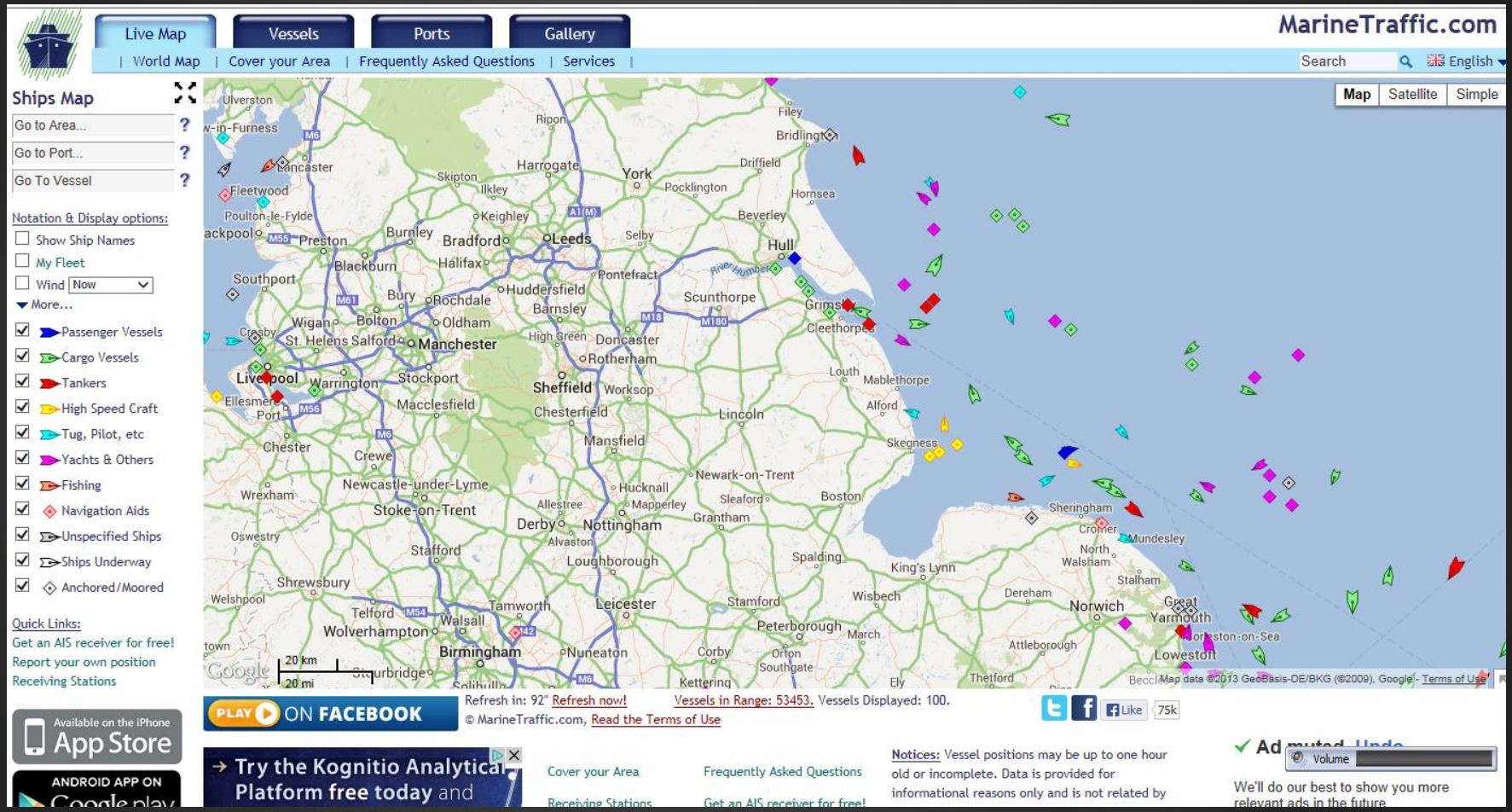


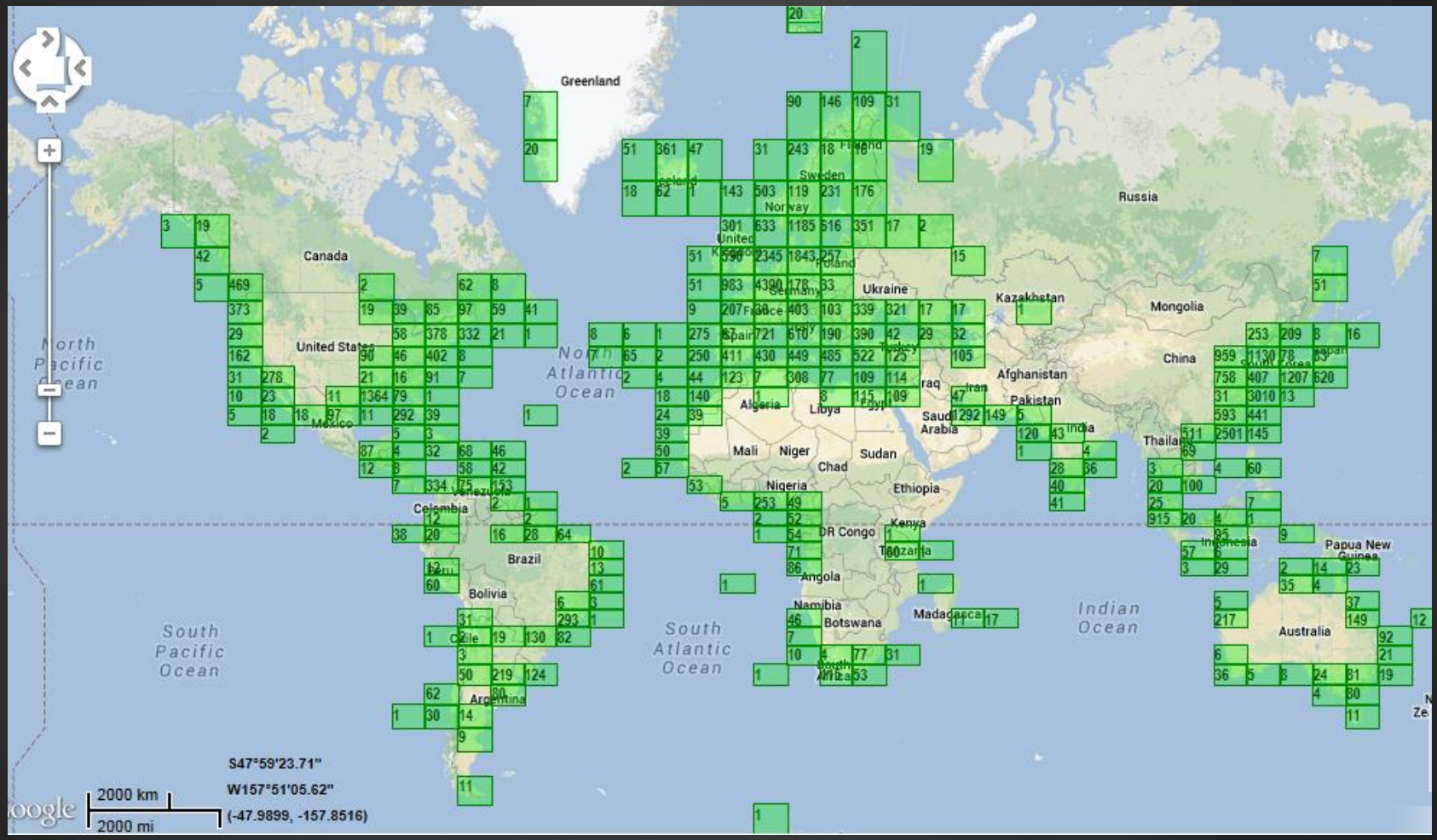


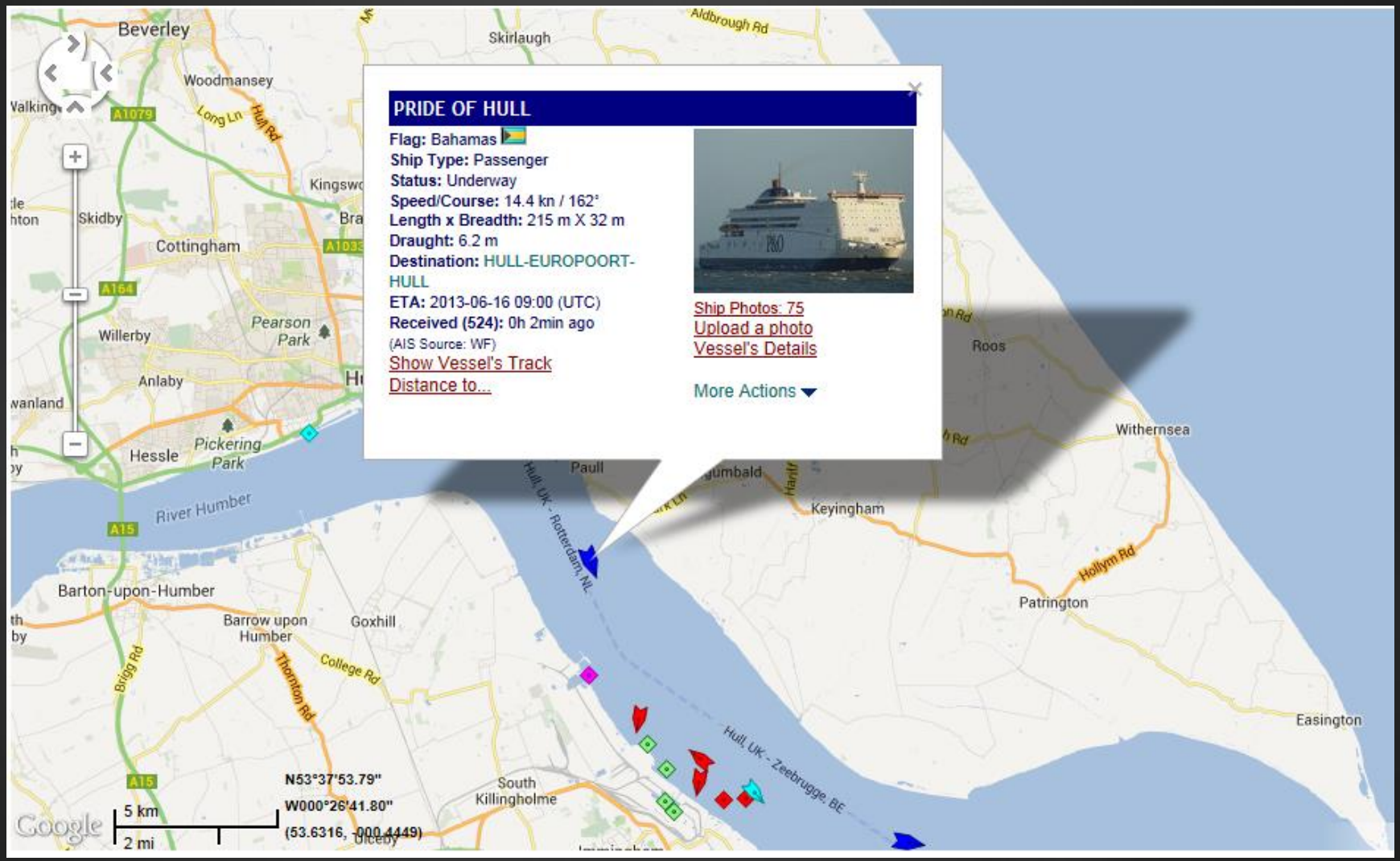


Understanding data requires different techniques and methodologies.

WHERE HAS BIG DATA BEEN APPLIED TO PRESENTLY?


Live Map
Vessels
Ports
Gallery
MarineTraffic.com
World Map | Cover your Area | Frequently Asked Questions | Services
Search
English
Ships Map
Go to Area...
Go to Port...
Go To Vessel
Notation & Display options:
Show Ship Names
My Fleet
Wind Now
More...
Passenger Vessels
Cargo Vessels
Tankers
High Speed Craft
Tug, Pilot, etc
Yachts & Others
Fishing
Navigation Aids
Unspecified Ships
Ships Underway
Anchored/Moored
Quick Links:
Get an AIS receiver for free!
Report your own position
Receiving Stations
Map
Satellite
Simple
Liverpool
Manchester
Leeds
Sheffield
Hull
Nottingham
Birmingham
Leicester
Norwich
Lincoln
Peterborough
Map data ©2013 GeoBasis-DE/BKG (©2009), Google - Terms of Use
PLAY ON FACEBOOK
Refresh in: 92" Refresh now!
Vessels in Range: 53453. Vessels Displayed: 100.
© MarineTraffic.com, Read the Terms of Use
Like 75k
Available on the iPhone App Store
ANDROID APP ON Google play
Try the Kognitio Analytical Platform free today and
Cover your Area
Frequently Asked Questions
Receiving Stations
Get an AIS receiver for free!
Notices: Vessel positions may be up to one hour old or incomplete. Data is provided for informational reasons only and is not related by
Volume
We'll do our best to show you more relevant ads in the future


Greenland
Canada
United States
Mexico
Russia
Kazakhstan
Mongolia
China
Afghanistan
Pakistan
India
Thailand
Ukraine
Sweden
Norway
Poland
Germany
France
Spain
Italy
Algeria
Libya
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DR Congo
Angola
Namibia
Botswana
Brazil
Bolivia
Colombia
Chile
Argentina
Papua New Guinea
Australia
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
S47°59'23.71"
W157°51'05.62"
(-47.9899, -157.8516)
2000 km
2000 mi

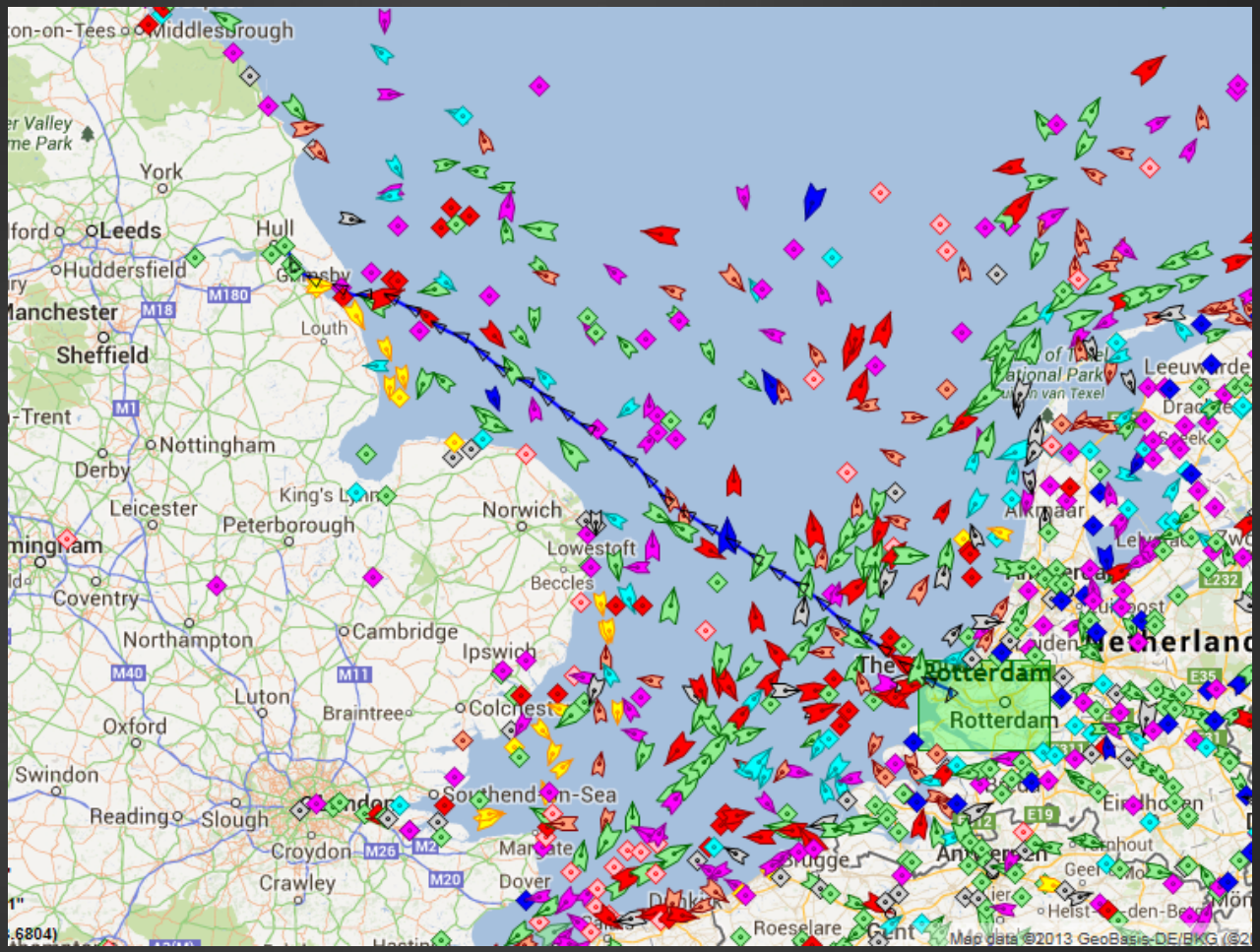

PRIDE OF HULL
Flag: Bahamas
Ship Type: Passenger
Status: Underway
Speed/Course: 14.4 kn / 162°
Length x Breadth: 215 m X 32 m
Draught: 6.2 m
Destination: HULL-EUROPOORT-HULL
ETA: 2013-06-16 09:00 (UTC)
Received (524): 0h 2min ago
(AIS Source: WF)
Show Vessel's Track
Distance to...
Ship Photos: 75
Upload a photo
Vessel's Details
More Actions
Beverley
Skirlaugh
Woodmansey
Cottingham
Skidby
Willerby
Anlaby
Hessle
Pearson Park
Pickering Park
River Humber
Barton-upon-Humber
Barrow upon Humber
Goxhill
South Killingholme
Paull
Roos
Withernsea
Keyingham
Patrington
Easington
Hull, UK - Rotterdam, NL
Hull, UK - Zeebrugge, BE
N53°37'53.79"
W000°26'41.80"
(53.6316, -000.4449)
5 km
2 mi
Google

PRIDE OF HULL

Flag: Bahamas
Ship Type: Passenger
Status: Underway
Speed/Course: 14.4 kn / 162°
Length x Breadth: 215 m X 32 m
Draught: 6.2 m
Destination: HULL-EUROPOORT-HULL
ETA: 2013-06-16 09:00 (UTC)
Received (524 : 0h 2min ago
(AIS Source: WF)
Show Vessel's Track
Distance to...

Ship Photos: 75
Upload a photo
Vessel's Details

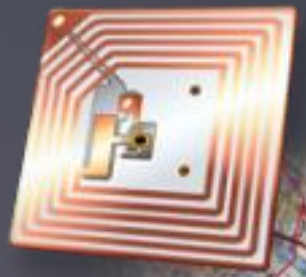
More Actions


Middlesbrough
York
Leeds
Hull
Huddersfield
Manchester
Sheffield
Louth
Nottingham
Derby
Leicester
Peterborough
Norwich
Lowestoft
Beccles
Coventry
Northampton
Cambridge
Ipswich
Luton
Braintree
Colchester
Oxford
Swindon
Reading
Slough
Croydon
Crawley
Dover
Rotterdam
Roeselare
Ghent
Map data ©2013 GeoBasis-DE/BKG

The fourth V = Visualisation

***12+ TBs* of tweet data every day**

30 billion RFID tags today (1.3B in 2005)



4.6 billion camera phones world wide

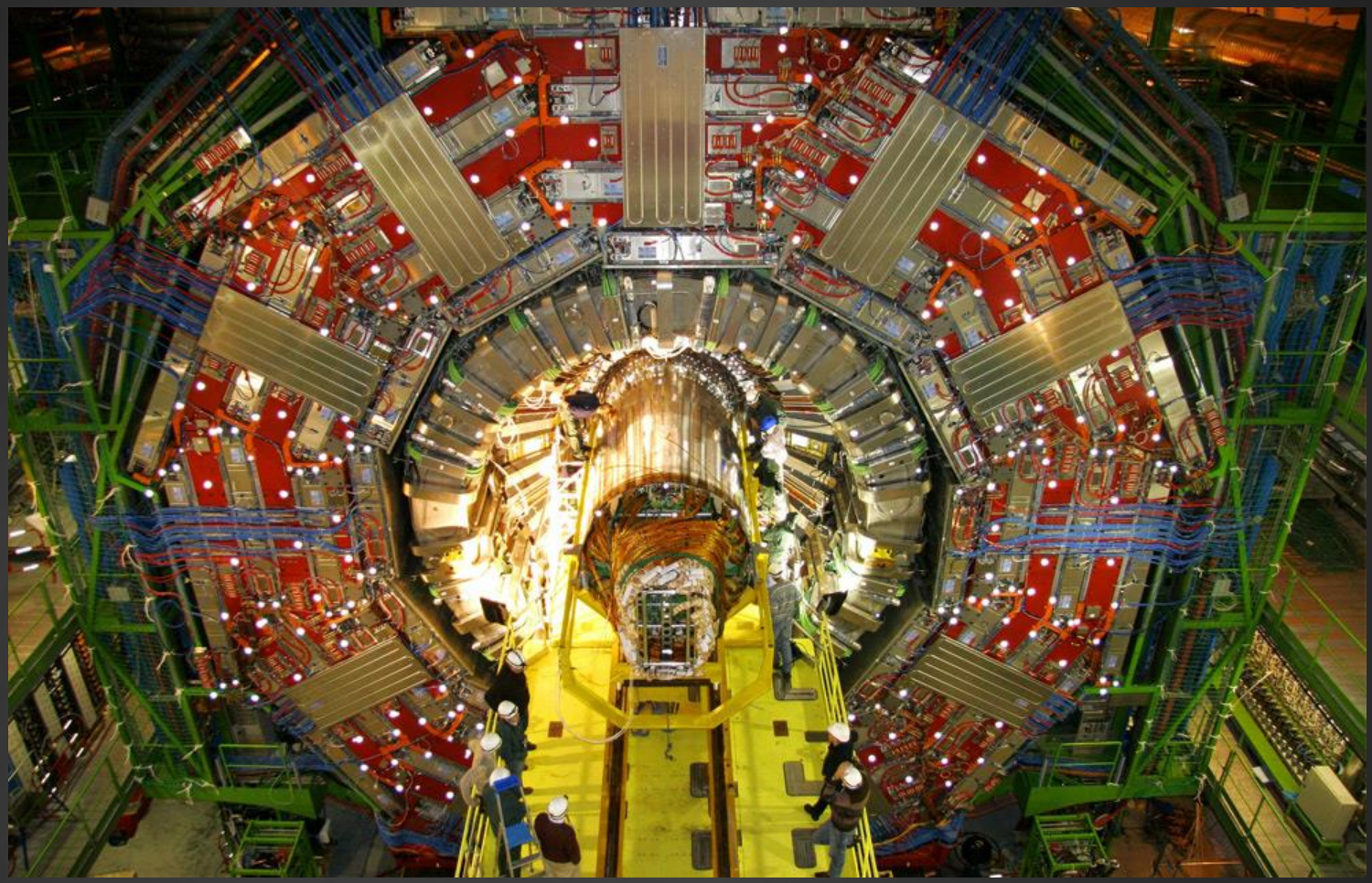
100s of millions of GPS enabled devices sold annually

2+ billion people on the Web by end 2011

76 million smart meters in 2009… 200M by 2014

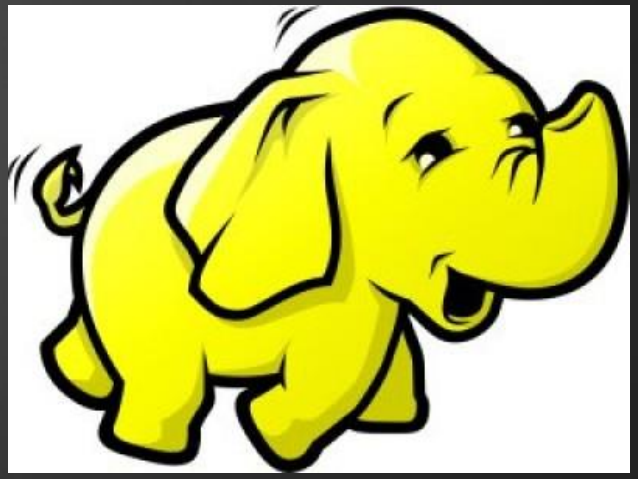
***2 TBs* of data every day**

***25+ TBs* of log data every day**



CERN's Large Hydron Collider (LHC) generates 15 PB a year

Misconceptions about Big Data is that the data is important, but what you do with it is even more important.



WHAT'S THE VALUE IN BIG DATA? CAN YOU THINK OF ANYTHING?

WHO IS USING IT?

Internet	Google, Amazon, eBay, AOL – everything
Mobile Gaming	On line betting, multi user games
Marketing	Social Network Analysis, Digital Advertising
Telcos	Call Routing Management, Subscriber Churn
Retail	Tesco – customer segmentation Netflix – everything
Utilities	Smart Grid for meters
Banking	Fraud detection, Customer segmentation
Insurance	Telemetrics on cars

WHERE IS IT APPLIED BEYOND THE OBVIOUS

- **The Automotive Industry**

According to this article, Ford's modern hybrid Fusion model generates up to 25 GB of data per hour. Why? The data can be used to understand driving behaviors and reduce accidents, understand wear and tear to identify issues that lower maintenance costs, avoid collisions, and even confirm travelling arrangements.

WHERE IS IT APPLIED BEYOND THE OBVIOUS

- **Supply Chain, Logistics, and Industrial Engineering**

Companies like Union Pacific Railroad use thermometers, microphones, and ultrasound to capture data about their engines and send it for analysis to identify equipment at risk for failure. INTTRA, the world's largest, multi-carrier network for the ocean shipping industry, uses it's OceanMetrics application to allow shippers and carriers to measure their own performance. As well, companies are using telematics and big data to streamline trucking fleets and how they can improve fuel usage and routes. GE believes these types new capabilities can contribute $15 trillion to the global GDP by 2030 by using systematic, data-driven analysis to trim costs and waste.

WHERE IS IT APPLIED BEYOND THE OBVIOUS

Retail

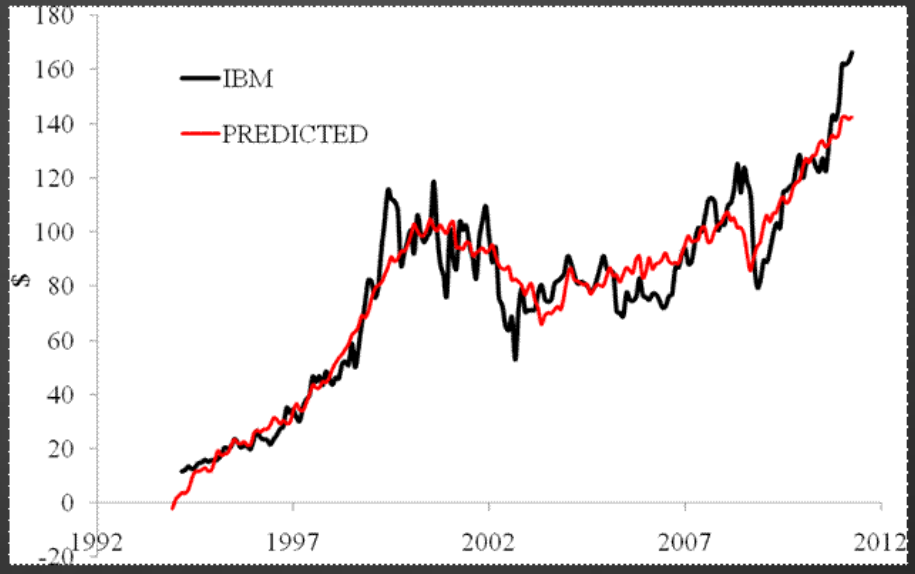
Walmart is using big data from 10 different websites to feed shopper and transaction data into an analytical system. Sears and Kmart are trying to improve the personalization of marketing campaigns, coupons, and offers with big data to compete better with Wal-Mart, Target, and Amazon. As the leader in the space, Amazon uses 1 million Hadoop clusters to support their affiliate network, risk management, machine learning, website updates, and more.

WHERE IS IT APPLIED BEYOND THE OBVIOUS

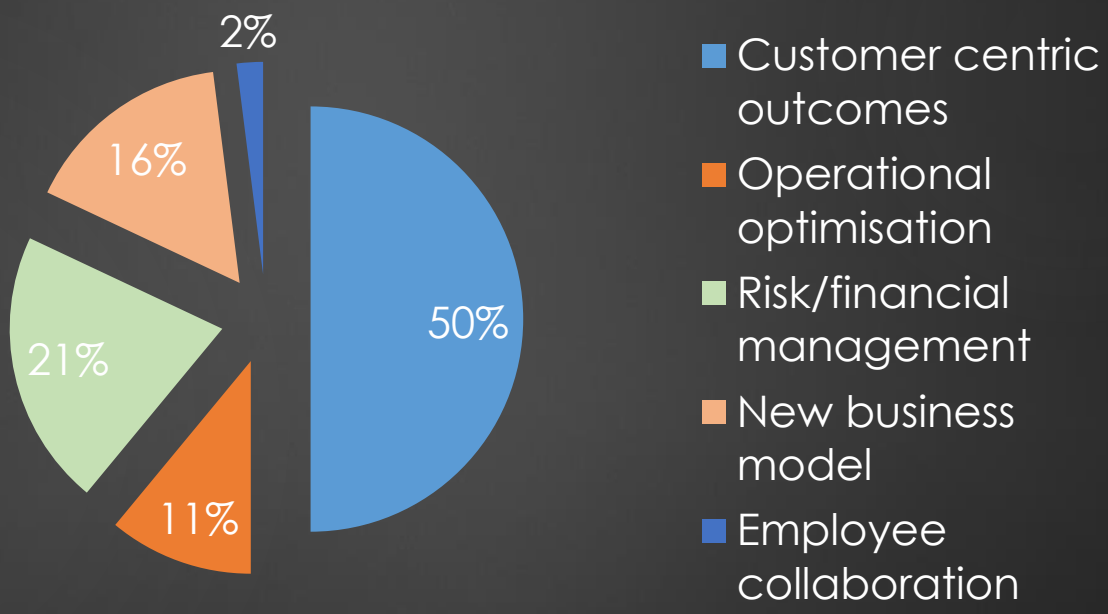
ENTERTAINMENT

Companies like Time Warner, Comcast, and Cablevision are using big data to track media consumption and engagement, advertising, and customer retention as well as operations and infrastructure. The video game industry is using big data for tracking during gameplay and after, predicting performance, and analyzing over 500GB of structured data and 4 TB of operational logs each day. Even brands like ESPN are looking to get in on the action.

BIG DATA CAN BE USED IN FINANCE, BUT IT IS NOT PREVALENT YET.



So how is Big Data used in finance? NOT MUCH TRADING YET.



Source: IBM 2012

So who uses big data in finance?

- Monitor customer journeys:
  - website clicks
  - transaction records
  - bankers' notes
  - voice recordings
- Make pre-emptive competitive credit offers
- Selective BankAmeriDeals cash back offers

So who uses big data in finance?

- Insurance uses:
- Fraud Detection & Analysis
- Personalised Pricing
- Customer Sentiment Analysis
- Catastrophic Planning
- Call Detail Record
- Loyalty Management
- Social Media Analytics
- Advertising and Campaign Management

AND LASTLY…

The Financial Services Industry

Of course, it's probably no surprise the financial services industry wants to use data to make better financial decisions. For example, Morgan Stanley ran into issues doing portfolio analysis on traditional databases and now uses Hadoop to analyze investments "on a larger scale, with better results." As well, Hadoop is being used in the industry for sentiment analysis, predictive analytics, and financial

Lastly, Big data in retail - Tesco

Major use of BI and loyalty program

Outsourced to Dunnhumby for analysis
80% of customers are Clubcard members

Constantly updating data – every transaction improves data quality

Targeted promotions both direct and sponsored by suppliers (revenue stream)

But Tesco is a financial services company as well...